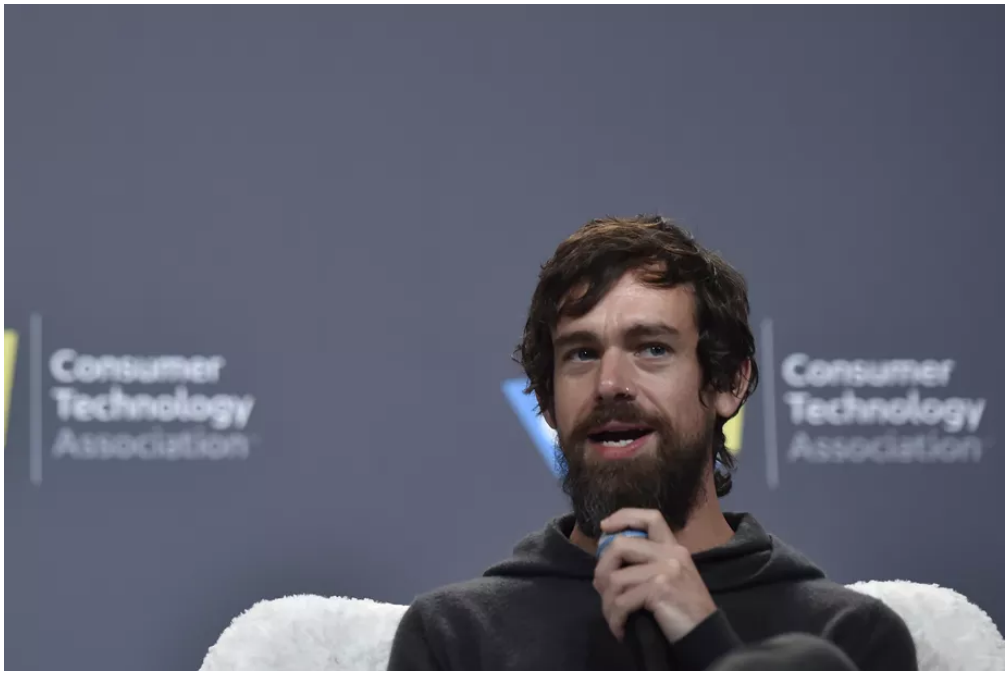
Exhibit 1

Twitter says it's getting better at detecting abusive tweets without your help

Twitter is using technology to catch more bad tweets.

By Kurt Wagner | Apr 16, 2019, 3:35pm EDT



Twitter CEO Jack Dorsey. | David Becker/Getty Images

Twitter can be a terrible, hateful place. It's why the company has promised over and over and over again that it plans to clean up its service and fight user abuse.

Part of the problem with that cleanup effort, though, has been that Twitter predominantly relies on its users to find abusive material. It wouldn't (or couldn't) find an abusive tweet without someone first flagging it for the company. With more than 300 million monthly users, that's a near-impossible way to police your service.

Good news: Twitter says it's getting better at finding and removing abusive content without anybody's help.

In a blog post published Tuesday, Twitter says that "38 [percent] of abusive content that's enforced is surfaced proactively to our teams for review instead of relying on reports from people on Twitter."

The company says this includes tweets that fell into a number of categories, including "abusive behavior, hateful conduct, encouraging self-harm, and threats, including those that may be violent."

A year ago, 0 percent of the tweets Twitter removed from these categories were identified proactively by the company.

The blog post included a number of other metrics Twitter shared to try and convey to people that Twitter is getting safer, but the 38 percent number was the most important. The reality of having a platform as large as Twitter's is that it is impossible to monitor with humans alone. This technology is not just useful — it's a necessity.

Facebook, for example, has for years been proactively flagging abusive posts with algorithms. With "hate speech," Facebook says last fall it removed more than 50 percent of posts using algorithms. In the "violence and graphic content" category, it proactively identified almost 97 percent of violating posts. For "bullying and harassment," Facebook is still just at 14 percent.

Algorithms are far from foolproof. On Monday, as video of the Notre Dame Cathedral burning was shared on YouTube, the company's algorithms started surfacing September 11 terrorist attack information alongside the videos, even though they are not related events. When a shooter opened fire at a New Zealand mosque late last month, algorithms on Facebook, YouTube, and Twitter couldn't stop the horrific videos from spreading far and wide.

But algorithms designed to improve safety are the only way Twitter is going to keep pace with the volume of tweets people share every day. Twitter is far from "healthy," but it may be getting a little closer to cleaning up its act.

One element missing from Twitter's blog: Any update on its efforts to actually *measure* the health of its service, something Twitter announced over a year ago it would work on. Those efforts have been slow, but Twitter executives told **Recode** last month that some

of their work in measuring the health of the service could appear in the actual product as early as this quarter.

Recode Daily

Email (required)

By signing up, you agree to our Privacy Notice and European users agree to the data transfer policy.

SUBSCRIBE

This article originally appeared on Recode.net.

Millions turn to Recode to understand how technology and the companies behind it are shaping our world — and what's at stake as we rely on technology more than ever before. Financial contributions from readers help support our journalism and enable our staff to continue to offer our articles, podcasts, and newsletters for free. Please consider making a contribution today from as little as $3.


.gif (1×3) - iframe-bottom

.gif (1×3) - iframe - doctem